A Modelling Environment for Mind and Matter Aspects of Intentional Behaviour

Catholijn M. Jonker, Jan Treur, Wouter C.A. Wijngaards
Department of Artificial Intelligence, Vrije Universiteit Amsterdam,
De Boelelaan 1081a, 1081 HV Amsterdam, The Netherlands
Email: <{jonker,treur,wouterw}@cs.vu.nl>
URL: http://www.cs.vu.nl/~{jonker,treur,wouterw}

Abstract. In this paper the internal dynamics of mental states, in particular states based on beliefs, desires and intentions, is formalised using a temporal language. A software environment is presented that can be used to specify, simulate and analyse temporal dependencies between mental states in relation to traces of them. If also relevant data on internal physical states over time are available, these can be analysed with respect to their relation to mental states.

1 Introduction

Dynamics has become an important focus within Cognitive Science in recent years; e.g., [12]. As one of the aspects, the dynamics of the interaction with the external world, and its implications for the representational content and dynamics of mental states have received attention; e.g., [2], [5]. Another important aspect is the internal dynamics of mental states, as can be found, for example in the dynamics of intentional notions (such as beliefs, desires and intentions) and their interaction with each other and with the external world. An example of a pattern for such internal dynamics is: if a desire and an additional reason (in the form of a belief about the world) to do some action are both present, then the intention to do the action is generated.

In this paper the internal dynamics of mental states, based on beliefs, desires and intentions (which also may include dynamics of the interaction of mental states with the external world) is addressed. A modelling environment is presented that can be used to specify, simulate and analyse models for these dynamics taking into account mental aspects (mind), physical aspects (matter), or both. A basic notion underlying the modelling is the notion of functional role or profile of a mental state, cf. [3].

A question is how such functional roles can be modelled in a precise and formal manner that stays close to the original idea. In this paper functional roles of belief, desire and intention states are modelled in a *temporal language* in such a manner that causal relationships are formalised by temporal dependencies they entail. Since dynamics is a phenomenon occurring over real time, the real numbers are used as time frame; no approximation by a sequence of fixed discrete time steps is needed. The temporal language can be used on the one hand for the *specification* of temporal relationships between mental states involving beliefs, desires and intentions (and between mental states and the external world). Such a temporal specification can be used to express a theory for these dynamics. On the other hand the language is the basis of a *software environment* that has been implemented and which can be used for the simulation and analysis of the internal dynamics.

Simulation takes place within this software environment by generating consequences over time from the specified set of temporal relationships, according to the paradigm of executable temporal logic [1]. To predict the internal dynamics, the software takes the temporal relationships, some initial values, and a pattern of environment dynamics to produce implied traces of internal belief desire and intention states. *Analysis* of given traces (in comparison to certain temporal relationships) is supported by the software environment as well. For example, these given traces can have the form of successively attributed intentional states over time. The automated support displays any discrepancies between such data and a background theory of the dynamics expressed by (assumed) temporal relationships. Another use of the software environment is the analysis of the relationship between mental and physical internal states. If observations (e.g., by advanced scanning techniques) can be made of the physical states assumed to be related to mental states, these empirical physical traces can be used as input, after which the software environment generates the related mental traces and checks the temporal relationships.

In Section 2 the intentional notions on which the paper focuses are introduced; for each type of intentional notion its functional role with respect to the other notions is discussed informally. In Section 3 the formalisation for the dynamics is presented. An example is discussed in Section 4. Subsequently in Section 5 the software environment, and some results are presented. Section 6 addresses the use of the environment when relevant physical internal state data over time are available, while Section 7 concludes with a discussion.

2 The Intentional Notions Addressed

The intentional notions from the BDI model (belief, desire and intention), are addressed in a static manner in e.g. [13], [11]; in our approach they are used in temporal perspective, see Figure 1. For an approach exploiting such a temporal perspective to attribute intentional notions on the basis of observed behaviour, see [10].

Beliefs are based on observation of the outside world in the present or in the past. Beliefs are modified in response to changes perceived in the external world. A belief denoted by the property $\beta(x, pos)$ means that the agent thinks that the property x holds. The property $\beta(x, neg)$ means that the agent thinks that the property x does not hold. Beliefs can be incorrect (a *false belief)*, e.g. due to some faulty sensory input.

It is possible to have both the belief that something holds and the belief that it does not hold, at the same time. Although such a state of affairs may have deplorable consequences for the agent, it is expressible.

Desires are states of the world or changes to the world that are desired. Desires are formed based on the agent's history. Desires are created and stay in existence for a while. The property $\delta(x)$ denotes a desire for x. The desire can be for a situation or an action. The desires the agent has at one time can conflict with each other.

From the set of desires that exist in a given situation some can be chosen to be pursued by creating an *intention* for them. For example, when a desire exists and an additional reason ρ (i.e. a particular co-occurrence of beliefs) also holds then an intention to fulfil the desire is created. This intention lasts until the desire or the additional reason for it disappears. *Additional reasons* perform at least two functions. Firstly, they inhibit the selection of conflicting intentions. Secondly, they cause the selection of particular intentions when those intentions are appropriate. The first and second uses can overlap. For example, if an animal obtains food, it could intend to eat it, or store it for later use. The intention to store the food for later, could need the reason that winter is approaching, selecting the intention when appropriate. The intention to store the food is used under the condition (additional reason) that it is not hungry, preventing a conflict with the intention to eat the food, which it only does when hungry.

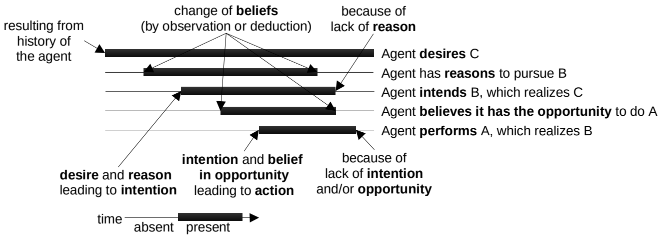
change of **beliefs**
(by observation or deduction)
because of
lack of **reason**
resulting from
history of
the agent
Agent **desires** C
Agent has **reasons** to pursue B
Agent **intends** B, which realizes C
Agent **believes it has the opportunity** to do A
Agent **performs** A, which realizes B
desire and **reason**
leading to **intention**
intention and **belief**
in opportunity
leading to **action**
because of
lack of **intention**
and/or **opportunity**
time
absent
present

Fig 1. BDI notions over time

The intentions are states or changes in the world that are intended to be accomplished. An intention is denoted $\iota(x)$. The intentions of an agent at a particular time do not conflict with each other. When the intention exists and it is believed that an *opportunity* o presents itself, the *action* is performed. The action is undertaken until the intention or the belief in the opportunity for it disappears. An action atom θ of the form $\alpha(x)$ refers to process x in the external world. Actions can have the intended effect, but can also fail or produce unexpected results.

3 Dynamical Formalisation

In BDI-logics such as [13], [11], internal processes are considered instantaneous. However, a more sincere formalisation is obtained if also internal processes take time. In this paper real time is used (represented by real numbers); time is not measured in computational steps. Real time temporal relationships are defined that take into account the delay between cause and effect, together with the durations of those cause and effect situations. The delay and durations may be measured. In this setting, the BDI-notions can be defined by the functional role they play. In the following the term *agent* is used to refer to the subject and *system* is used to refer to both the agent and the external world together. Intervals of real numbers are denoted like: $[x, y)$ meaning $\{p \in \mathbb{R} \mid p \geq x \wedge p < y\}$. Thus, '[' or ']' stands for a closed end of the interval, and '(' or ')' stands for an open end of the interval.

Definition (state properties)

The states of the system are characterised by *state properties*. The state properties are formalised using (logical) formulae over a specific ontology. For an ontology Ont, the set of *atoms* AT(Ont) contains the atomic properties expressed in terms of the ontology. The set of *state properties* SPROP(Ont) contains all the propositional formulas built out of the atoms using standard propositional connectives. More specifically, the following ontologies are used. Firstly, *world state properties* express properties of a particular situation in the material world, using ontology EWOnt. Secondly, the internal physical state properties of the agent are expressed using IntOntP. The combined physical ontology is $\mathrm{OntP} =_{\mathrm{def}} \mathrm{EWOnt} \cup \mathrm{IntOntP}$. Thirdly, the ontology for internal mental state properties is denoted by IntOntM. The ontology for all state properties is denoted by $\mathrm{AllOnt} =_{\mathrm{def}} \mathrm{EWOnt} \cup \mathrm{IntOntP} \cup \mathrm{IntOntM}$.

Definition (states)

a) A *physical state* P of the system is an assignment of truth values {true, false} to the set of physical state atoms AT(OntP) of the system. The set of all possible physical states is denoted PS.
b) A (partial) *mental state* M of the system is an assignment of truth values {true, false, unknown} to the set of internal mental state atoms, AT(IntOntM). The set of all possible mental states is denoted by MS. Three valued states are used to avoid commitment to closed world assumptions or explicit specification of negative conclusions.
c) At each time-point the system is in one state. This state is from the set $\mathrm{States} =_{\mathrm{def}} \mathrm{PS} \times \mathrm{MS}$.
d) The standard satisfaction relation $\models$ between states and state properties is used: $S \models \varphi$ means that property φ holds in state S.

Definition (traces)

The system when viewed over a period of time, will produce several states consecutively. The function $\mathcal{T}$ returning the state for each time point is called a trace, $\mathcal{T}: \mathbb{R} \to \mathrm{States}$. The notation $\mathrm{state}(\mathcal{T}, t, m)$, where $\mathcal{T}$ is a trace, $t \in \mathbb{R}$ and $m \in \{\mathrm{physical}, \mathrm{mental}\}$, means the physical or mental state at time t in trace $\mathcal{T}$. The notation $\mathrm{state}(\mathcal{T}, t)$ is by definition $\mathcal{T}(t)$. Thus using the last notation both physical and mental terms can be used interchangeably, under the assumption that $\mathrm{PS} \cap \mathrm{MS} = \varnothing$. The set of all possibly occurring traces is denoted $\mathcal{W}$.

The behaviour of the agent and environment is defined by a set of traces. Temporal relationships between the state properties over time specify such a set of traces: they express certain constraints on the relative timing of the occurrence of state properties. These constraints on the timing reflect a causal relationship between the arguments.

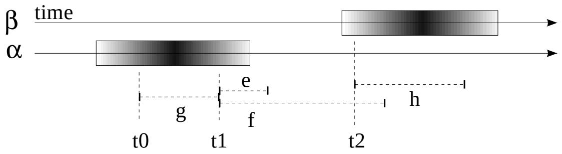

Fig 2. The time relationships between variables

Definition (the '$\twoheadrightarrow$' relation and the '$\bullet\!\!-$' relation)

Let α , $\beta \in \mathrm{SPROP(AllOnt)}$. The state property α *follows* state property β, denoted by $\alpha \twoheadrightarrow_{e,\,f,\,g,\,h} \beta$, with time delay interval $[e, f]$ and duration parameters g and h if
$\forall \mathcal{T} \in \mathcal{W}\ \forall t1$:

$$[\forall t \in [t1 - g, t1) : \mathrm{state}(\mathcal{T}, t) \models \alpha \ \Rightarrow\ \exists d \in [e, f]\ \forall t \in [t1 + d, t1 + d + h) : \mathrm{state}(\mathcal{T}, t) \models \beta\ ]$$

Conversely, the state property β *originates from* state property α, denoted by $\alpha \bullet\!\!-_{e,\,f,\,g,\,h} \beta$, with time delay in $[e, f]$ and duration parameters g and h if

$\forall\ \mathcal{T} \in \mathcal{W}\ \forall\ \mathrm{t2}$:

$[\forall \mathrm{t} \in [\mathrm{t2},\ \mathrm{t2} + \mathrm{h})\ :\ \mathrm{state}(\mathcal{T}, \mathrm{t}) \models \beta \Rightarrow\ \exists \mathrm{d} \in [\mathrm{e},\ \mathrm{f}]\ \forall \mathrm{t} \in [\mathrm{t2} - \mathrm{d} - \mathrm{g},\ \mathrm{t2} - \mathrm{d})\ \ \mathrm{state}(\mathcal{T}, \mathrm{t}) \models\ \alpha]$

If both $\alpha \twoheadrightarrow_{e,f,g,h} \beta$, and $\alpha \bullet\!\!- _{e,f,g,h} \beta$ hold, this is denoted by: $\alpha \bullet\!\!\twoheadrightarrow_{e,f,g,h} \beta$.

The relationships between the variables α, β, e, f, g, h, t0, t1 and t2 are depicted in Figure 2.

Definition (continuous)

Let $\varphi, \psi \in \mathrm{SPROP(AllOnt)}$. The relationship $\varphi \bullet\!\!\twoheadrightarrow_{e,f,g,h} \psi$ is continuous if:
$\forall \mathcal{T} \in \mathcal{W} \forall \mathrm{t0}\ \forall \mathrm{t1}{>}\mathrm{t0}$: if $(\forall \mathrm{t} \in [\mathrm{t0},\ \mathrm{t1})\ :\ \mathrm{state}(\mathcal{T}, \mathrm{t}) \models \varphi)$ then
$(\forall \mathrm{t2},\mathrm{t3} \in\ [\mathrm{t0{+}g{+}e},\ \mathrm{t1{+}f{+}h}]$: $\mathrm{state}(\mathcal{T}, \mathrm{t2}) \models \psi\ \wedge\ \mathrm{state}(\mathcal{T}, \mathrm{t3}) \models \psi\ )$
$\Rightarrow (\forall \mathrm{t4} \in\ (\mathrm{t2},\ \mathrm{t3})$: $\mathrm{state}(\mathcal{T}, \mathrm{t4}) \models \psi)$.

Loosely phrased, continuous means that when φ holds for a continued length of time, then ψ will also hold for a continued length of time, without gaps.

Lemma

a) If $\mathcal{T}$ is a given trace, $\varphi, \psi \in \mathrm{SPROP(AllOnt)}$, $\varphi \twoheadrightarrow_{e,f,g,h} \psi$ and
$(\forall \mathrm{t} \in [\mathrm{t0},\ \mathrm{t0{+}g})\ :\ \mathrm{state}(\mathcal{T}, \mathrm{t}) \models \varphi)$ then a *guaranteed result* exists:
$(\forall \mathrm{t} \in\ [\mathrm{t0{+}g{+}f},\ \mathrm{t0{+}g{+}e{+}h})\ :\ \mathrm{state}(\mathcal{T}, \mathrm{t}) \models \psi)$.
b) If $\mathcal{T}$ is a given trace, $\varphi, \psi \in \mathrm{SPROP(AllOnt)}$, $\varphi \bullet\!\!- _{e,f,g,h} \psi$ and
$(\forall \mathrm{t} \in [\mathrm{t0},\ \mathrm{t0{+}h})\ :\ \mathrm{state}(\mathcal{T}, \mathrm{t}) \models \psi)$ then a *guaranteed precondition* exists:
$(\forall \mathrm{t} \in\ [\mathrm{t0{-}e{-}g},\ \mathrm{t0{-}f})\ :\ \mathrm{state}(\mathcal{T}, \mathrm{t}) \models \varphi)$.

Proposition

If $\varphi \bullet\!\!\twoheadrightarrow_{e,f,g,h} \psi$ and $\mathrm{e} + \mathrm{h} \geq \mathrm{f}$, then $\varphi \bullet\!\!\twoheadrightarrow_{e,f,g,h} \psi$ is continuous.

An interval of ψ of $[\mathrm{t0{+}g{+}f},\ \mathrm{t1{+}e{+}h})$ can be seen to hold continuously when given an interval of φ of $[\mathrm{t0},\ \mathrm{t1})$, using the lemma of guaranteed result, to assure there are no gaps in between. In order for the result to keep holding, when the antecedent keeps holding, the parameters of $\bullet\!\!\twoheadrightarrow$ should have certain values. If $\mathrm{e} + \mathrm{h} \geq \mathrm{f}$ then for each application of the definition of the relation we can be sure that the period $[\mathrm{t1} + \mathrm{f},\ \mathrm{t1} + \mathrm{e} + \mathrm{h}]$ holds. To see how this can be, consider that a range of resulting intervals is possible, with at the earliest $[\mathrm{t1} + \mathrm{e},\ \mathrm{t1} + \mathrm{e} + \mathrm{h}]$ and at the last $[\mathrm{t1} + \mathrm{f},\ \mathrm{t1} + \mathrm{f} + \mathrm{h}]$. With $\mathrm{e} + \mathrm{h} \geq \mathrm{f}$ holding, the two intervals will overlap, this overlap is exactly the interval $[\mathrm{t1} + \mathrm{f},\ \mathrm{t1} + \mathrm{e} + \mathrm{h}]$. Thus if $\mathrm{e} + \mathrm{h} \geq \mathrm{f}$ and the φ holds in a long interval $[\mathrm{t3},\ \mathrm{t4}]$, where $\mathrm{t4} - \mathrm{t3} \geq \mathrm{g}$ then ψ will hold in the interval $[\mathrm{t3} + \mathrm{f} + \mathrm{g},\ \mathrm{t4} + \mathrm{e} + \mathrm{h}]$.

Definition (internal belief representation)

Let $\varphi \in \mathrm{SPROP(OntP)}$ be a physical state property.
a) The internal mental state property $\beta \in \mathrm{SPROP(IntOntM)}$ is called an *internal belief representation* for φ with time delay e and duration parameters f, g if: $\varphi\ \bullet\!\!\twoheadrightarrow_{e,f,g,h} \beta$.

b) Two belief representations β_1 and β_2 are *exclusive* if
$\forall \mathcal{T} \in \mathcal{W}: \neg\exists \mathrm{t}$: $\mathrm{state}(\mathcal{T}, \mathrm{t}) \models \beta_1 \wedge \beta_2$.

In a) of this definition the $\twoheadrightarrow$ part is necessary, as the occurrence of external state φ should lead to the creation of the belief β. The $\bullet\!\!-$ part must also hold, since a belief

β must have an explanation of having being created, in this case φ. This consideration also holds for intentions and desires in an analogical fashion.

When the world situation suddenly changes, the beliefs may follow suit. The belief β_1 and the belief β_2 of two opposite world properties should not hold at the same time; they should be exclusive. As the external world state fluctuates, the beliefs should change accordingly, but never should there be both a belief for a world property and a belief for the opposite world property at the same time. If two belief representations for opposite world properties are exclusive, this inconsistency is avoided, and the belief representations are called *non-conflicting*.

Definition (internal intention representation)

Let $\alpha \in$ SPROP(OntP) be a physical state property, $\beta \in$ SPROP(IntOntM) a belief representation for α and $\theta \in$ SPROP(IntOntM) an action atom. The internal mental state property $\gamma \in$ SPROP(IntOntM) is called an *internal* i*ntention representation* for action θ and opportunity α with delay e, f and duration parameters g, h if $\gamma \wedge \beta \bullet\!\!\rightarrow\!\!\!\rightarrow_{e,f,g,h} \theta$.

Definition (internal desire representation)

Let $\rho \in$ SPROP(OntP) be a physical state property, β a belief representation for ρ and γ an intention representation. The internal mental state property $\delta \in$ SPROP(IntOntM) is an *internal desire representation* for intention γ and additional reason ρ with delay e, f and duration parameters g, h if $\delta \wedge \beta \bullet\!\!\rightarrow\!\!\!\rightarrow_{e,f,g,h} \gamma$.

4 An Example Formalisation

In order to demonstrate the formalisation and automated support put forward in this paper, a simple example description is presented. In this example, the test subject is a common laboratory mouse, that is presented with cheese. Mostly, the mouse will try to eat the cheese, but a transparent screen can block access to the cheese. First, an intentional perspective on the mouse is constructed. Then, assuming a mouse-brain-scanning-technique, it is analysed how specific brain area activity can be correlated to the intentional notions.

The formalised physical external world description of this experiment has two properties; screen_present and cheese_present. The internal physical state has the property hungry. The intentional description of the mouse makes use of the following beliefs on the relevant parts of the world for this experiment: β(hungry, pos), β(hungry, neg), β(screen_present, pos), β(screen_present, neg), β(cheese_present, pos) and β(cheese_present, neg). These beliefs are all based on perceptions by the mouse.

The beliefs should persist continuously if the perceptions stay the same. So if φ holds in the interval [t0, t2) then the belief will hold in a continuous resultant interval. The timing parameters of the belief observations indeed guarantee that a continuous belief representation is obtained.

When the world situation changes, the beliefs change. The g and h of the belief generation relations are chosen equal, so that the belief representations become double-seamless: the belief in a world property starts to be there exactly at the same time the belief in the opposite property stops to be there. By fixing the delays, as done in the double-seamless settings, the belief representations are non-conflicting.

Furthermore, the intentional description includes desires. If the mouse is hungry, it desires to eat, δ(eat_food). When sufficient additional reason, ρ_1, is present – the belief that there is cheese – the mouse will intend to eat the cheese, ι(eat_cheese). When the mouse believes that the opportunity, o_1, presents itself, the screen not being present, the mouse will eat the cheese, the action denoted by α(eat_cheese).

The temporal relationships for the intentional description of the mouse are given below. All e, f, g and h values for the temporal relationships are given in sequence, after the $\bullet\!\!\twoheadrightarrow$ symbol, in a certain time unit (e.g., 0.1 second). In this example only the positive and negative beliefs need to be present, in general however, the predicates of the properties could contain numbers to describe the world in more detail.

.... **Sensing** ....

hungry $\bullet\!\!\twoheadrightarrow_{1, 5, 10, 10}$ β(hungry, pos) ∧ ¬β(hungry, neg).

¬hungry $\bullet\!\!\twoheadrightarrow_{1, 5, 10, 10}$ β(hungry, neg) ∧ ¬β(hungry, pos).

cheese_present $\bullet\!\!\twoheadrightarrow_{1, 5, 10, 10}$ β(cheese_present, pos) ∧ ¬β(cheese_present, neg).

¬cheese_present $\bullet\!\!\twoheadrightarrow_{1, 5, 10, 10}$ β(cheese_present, neg) ∧ ¬β(cheese_present, pos).

screen_present $\bullet\!\!\twoheadrightarrow_{1, 5, 10, 10}$ β(screen_present, pos) ∧ ¬β(screen_present, neg).

¬screen_present $\bullet\!\!\twoheadrightarrow_{1, 5, 10, 10}$ β(screen_present, neg) ∧ ¬β(screen_present, pos).

.... **Internal Processes** ....

β(hungry, pos) $\bullet\!\!\twoheadrightarrow_{1, 5, 10, 10}$ δ(eat_food).

δ(eat_food) ∧ ρ_1 $\bullet\!\!\twoheadrightarrow_{1, 5, 10, 10}$ ι(eat_cheese).

ι(eat_cheese) ∧ o_1 $\bullet\!\!\twoheadrightarrow_{1, 5, 10, 10}$ α(eat_cheese).

ρ_1 = β(cheese_present, pos).

o_1 = β(screen_present, neg).

.... **World Processes** ....

α(eat_cheese) ∧ cheese_present $\bullet\!\!\twoheadrightarrow_{1, 5, 10, 10}$ ¬hungry.

In order to derive the converse of the previous temporal relationships, a temporal variant of Clark's completion is used [7].

¬β(hungry, pos) $\bullet\!\!\twoheadrightarrow_{1, 5, 10, 10}$ ¬δ(eat_food).

¬(δ(eat_food) ∧ ρ_1) $\bullet\!\!\twoheadrightarrow_{1, 5, 10, 10}$ ¬ι(eat_cheese).

¬(ι(eat_cheese) ∧ o_1) $\bullet\!\!\twoheadrightarrow_{1, 5, 10, 10}$ ¬α(eat_cheese).

¬(α(eat_cheese) ∧ cheese_present) $\bullet\!\!\twoheadrightarrow_{1, 5, 10, 10}$ hungry.

Also, at the start of derivation the intentional notions will be false, in particular the mouse initially does not believe anything. The starting value of each property is given for e + λ(f-e) + g time units.

5 Implementation of the Software Environment

A software environment has been made which implements the temporal formalisation of the internal dynamic behaviour of the agent. Following the paradigm of executable temporal logic, cf. [1], a 2700 line simulation program was written in C++ to automatically generate the consequences of the temporal relationships. The program is a special purpose tool to derive the results reasoning forwards in time, as in executable temporal logic.

The graph in Figure 3 shows the reaction of the mouse to changes in the environment. Time is on the horizontal axis. The world state properties and the intentional notions are listed on the vertical axis. The parameter λ is fixed at 0.25. A dark box on top of the line indicates the notion is true, and a lighter box below the line indicates that the notion is false.

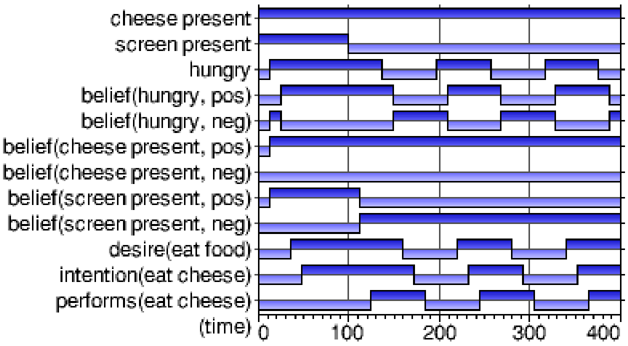

Fig 3. Simulation results: initially cheese and screen are present; later the screen is removed

As can be seen, the mouse is not hungry at the very start, but quickly becomes hungry. It desires to eat the cheese, and intends to do so, but the screen blocks the opportunity to do so. When the screen is removed, the mouse eats. After a while it stops eating, as it is not hungry anymore. Subsequently it enters a cycle where it becomes hungry, eats, and becomes hungry again. Intention revision is handled here by the leads-to operator and its duration parameters. After the consequent duration has passed, the intention is no longer required to hold. For example see Figure 3, where the intention to eat cheese ceases to hold, shortly after the desire to eat stops to hold.

Another program, of about 4000 lines in C++, has been constructed that takes an existing trace of behaviour as input and creates an interpretation of what happens in this trace and a check whether all temporal relationships hold. The program is configured (amongst others) by giving a set of intentional temporal relationships, see Section 4 for example relationships. The program marks any deficiencies in the trace compared with what should be there due to the temporal relationships. If a relationship does not hold completely, this is marked by the program. The program produces yellow marks for unexpected events. At these moments, the event is not produced by any temporal relationship; the event cannot be explained. The red marks indicate that an event has not happened, that should have happened. In addition to checking whether the rules hold, the checker produces an informal reading of the trace. The reading is automatically generated, using a simple substitution, from the information in the intentional trace.

6 Mind and Matter: Relating Physical and Mental States

In the formalisation, each internal state has a mental state and a physical state portion. The physical state is described by a set of (real number) value assignments to continuous variables. The automated support also supports the assignment of internal physical properties to intentional notions; also material data can be used as input. For the assignment of physical properties to intentions, each intentional property has one physical property associated. The values true and false of the intentional notion are assigned to particular ranges of values of the material in the data.

For the example, it is assumed that a scanner provides signal intensities for different brain areas, for modern scanning usage see [4]. Some of these may correlate with the intentions as described above. An assumed example assignment of intentional notions to the signal intensities of specific brain areas is given in Table 1.

The simulation program, having derived an intentional trace, can output a physical trace based on it. The physical trace consists of the possible ranges of values for all physical state variables in each time-interval.

The checker can read back a physical trace as generated by the simulator, but it can also read back a trace where for time-points a value for each physical state variable is given. It will then interpret this physical trace, comparing the given (range of) value(s) to the true and false ranges as given per intentional notion. It will then check whether all the given temporal relationships hold correctly.

Table 1. Related physical and mental state properties.

Intentional notion in SPROP(IntOntM)	Physical condition in SPROP(IntOntP)
β(hungry, pos)	intensity of area_01 ≥ 1.0
β(hungry, neg)	intensity of area_02 < 1.0
β(cheese_present, pos)	intensity of area_03 ≥ 1.0
β(cheese_present, neg)	intensity of area_04 < 1.0
β(screen_present, pos)	intensity of area_05 ≥ 1.0
β(screen_present, neg)	intensity of area_06 < 1.0
δ(eat_food)	intensity of area_07 ≥ 1.0
ι(eat_cheese)	intensity of area_08 ≥ 1.0
α(eat_cheese)	intensity of area_09 ≥ 1.0

Using the interpretation and checking of relationships the program can assist in the verification of hypothetical assignments of physical properties to intentional notions, and the verification of hypothetical intentional temporal relationships.

7 Discussion

This paper addresses formalisation of the internal dynamics of mental states involving beliefs, desires and intentions. In available literature on formalisation of intentional behaviour, such as [13], [11], the internal dynamics of intentional mental states are ignored. The formalisation of the internal dynamics of mental states introduced in this paper is based on a real time temporal language. Within this (quite expressive) temporal language a specific format is defined which can be used to specify temporal relationships that describe (constraints on) the dynamics of mental states (and their interaction with the external world). Specifications in this specific format have the advantage that they can be used to perform simulation, based on the paradigm of executable temporal logic, [1]. The approach subsumes discrete simulation, for example as performed in Dynamical Systems Theory [12] as a special case (with $e=f=1$ and $g=h=0$).

A software environment has been implemented including three programs. The first simulates the consequences of a set of temporal relationships of mental states over time. The second program interprets a given trace of intentional states over time (in terms of beliefs, desires and intentions), and makes an analysis whether the temporal relationships hold, and, if not, points at the discrepancies. A third program takes into account physical states and their (possible) relation to beliefs, desires and intentions. Physical traces, for example obtained by advanced scanning techniques [4, pp. 59-105], can be input and analysed with respect to possible interpretations in terms of mental properties such as beliefs, desires and intentions.

An example has been presented and explained: the internal dynamics of intentional eating behaviour of a mouse that in an experimental setting has to deal with a screen and cheese. The formalisation and supporting software environment is useful for simulation of the internal dynamics of mental states. In addition, they are useful for checking the attribution of intentions, (e.g., [8]) and predicting behaviour based on an attribution of intentions. Another use is if advanced scanning techniques provide empirical data. These data can be related to mental states and checked on correctness with respect to the required dynamics. The checking program can easily be used to check various assignments, and, for example, the number of bad marks per assignment. In this manner an assignment of materials to intentional notions can be selected from a set of hypothetically possible assignments.

In further research the attribution of intentional notions to explain the behaviour of some of the simpler biological organisms is addressed. Some first results have shown that the nutrient import behaviour, and the dynamics and control of the internal metabolism of the bacterium *E. Coli* can be explained using these notions.

References

1. Barringer, H., M. Fisher, D. Gabbay, R. Owens, & M. Reynolds (1996). *The Imperative Future: Principles of Executable Temporal Logic*, Research Studies Press Ltd. and John Wiley & Sons.
2. Bickhard, M.H. (1993). Representational Content in Humans and Machines. *Journal of Experimental and Theoretical Artificial Intelligence*, 5, pp. 285-333.
3. Bickle, J. (1998). *Psychoneural Reduction: The New Wave*. MIT Press, Cambridge, Massachusetts.
4. Chen, L. & Zhuo, Y., eds. (2001). *Proceedings of the Third International Conference on Cognitive Science (ICCS'2001)*. Press of University of Science and Technology of China.
5. Christensen, W.D. & C.A. Hooker (2000). *Representation and the Meaning of Life*. In: [5].
6. Clapin, H., Staines, P. & Slezak, P. (2000). *Proc. of the Int. Conference on Representation in Mind: New Theories of Mental Representation*, 27-29th June 2000, University of Sydney. To be published by Elsevier.
7. Clark, K.L. (1978). Negation as Failure. *Logic and Data Bases*. Gallaire, H. & Minker, J. (eds), Plenum Press, New York, pp. 293-322.
8. Dennett, D.C. (1987). *The Intentional Stance*. MIT Press. Cambridge Massachusetts.
9. Hodges, W. (1993). *Model Theory*. Cambridge University Press.
10. Jonker, C.M., Treur, J. & de Vries, W. (2001). "Temporal Requirements for Anticipatory Reasoning about Intentional Dynamics in Social Contexts", in: Y. Demazeau and F.J. Garijo (eds.), *Proceedings of MAAMAW'2001 (Modelling Autonomous Agents in a Multi-Agent World),* to appear in LNCS.
11. Linder, B. van, Hoek, W. van der & Meyer, J.-J. Ch. (1996). How to motivate your agents: on making promises that you can keep. In: Wooldridge, M.J., Müller, J. & Tambe, M. (eds.), *Intelligent Agents II. Proc. ATAL'95* (pp. 17-32). Lecture Notes in AI, vol. 1037, Springer Verlag.
12. Port, R.F. & Gelder, T. van (eds.) (1995). *Mind as Motion: Explorations in the Dynamics of Cognition*. MIT Press, Cambridge, Massachusetts.
13. Rao, A.S. & Georgeff, M.P. (1991). Modelling Rational Agents within a BDI-architecture. In: (Allen, J., Fikes, R. & Sandewall, E. ed.), *Proceedings of the Second International Conference on Principles of Knowledge Representation and Reasoning*, (KR'91), Morgan Kaufmann, pp. 473-484.